The Digital Shoebox

Serge Abiteboul

▶ **To cite this version:**

Serge Abiteboul. The Digital Shoebox. Memory, edited by Philippe Tortell, Mark Turin, and Margot Young, UBC Press, 2018. ⟨hal-01875161⟩

HAL Id: hal-01875161

https://inria.hal.science/hal-01875161v1

Submitted on 17 Sep 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

HAL Authorization

The Digital Shoebox

Serge Abiteboul

Inria & ENS, Paris

The digital information of our modern world — text, images, videos and the like, can be stored and reproduced massively at almost zero cost. It can be easily dispersed in space to preserve it from both the natural world (fire and water) and the political one (tyrants and censors). So why does it seem so difficult to preserve our digital memory?

At first blush, we might think the main problem results from changing digital formats. We have videos of our young children on VHS tapes, for example, which are no longer compatible with currently available hardware. This problem is certainly real, and exacerbated by rapid technological change; current information storage formats have relatively short lifetimes (from a few years to perhaps a few decades at most). Clearly, these recording formats are more ephemeral than Sumerian tablets, or paper. Yet, relatively simple solutions exist, using programs that translate and transcode between old and new formats, replicating information to guarantee the existence of complete and compatible copies. We are not always aware of these solutions, and they require effort (sometimes significant) on our part. Nonetheless, the preservation of digital information is possible, and typically less expensive than the reproduction and storage of physical records.

So where is the problem? It lies, fundamentally, in the deluge of data that forces us to choose *what we want to preserve and what we are willing to forget*. It is simply not possible to keep everything, nor do we necessarily want to. In particular, we forget details to gain insight, to be able to abstract. Herein lies the existential problem of digital memory: the choice of what to forget.

To illustrate this problem, let us consider photography. In the early days of this art form, when the production of a picture was expensive, people used their cameras sparingly, focusing their efforts on a small number of carefully chosen subjects. As the price of photographic reproduction decreased, however, output increased and the world became flooded with images. This became particularly true with the advent of digital photography. Today, a person can easily take several hundred or even thousand photos each year, mostly in a digital format that can be easily shared around the globe. In 2016, for example, 95 million photos and videos were posted every day on Instagram. Viewing all of these images, whether in a personal collection, or in a globally distributed platform, is becoming an increasing challenge, requiring significantly more time than many will be ready to spend.

Historically, we used to sort our information. Perhaps we had a shoebox where our most precious pictures were kept. The most organized of us were making albums. Today, where are our pictures? Somewhere on an Instagram or Facebook account, perhaps, or on our phone, a computer, or on an external drive. Typically, our digital information is spread across a combination of platforms, and on devices, any one of which can be damaged, stolen or

hacked. Digital cloud storage simplifies our lives by avoiding the dependency on specific hardware we must manage ourselves. But we can also get lost in the clouds, and a provider may decide, unbeknownst to us, not to archive our data beyond a few years. We change computers, we close accounts, time passes, and we lose entire portions of our memory.

What can we do? Perhaps we can create 'digital shoeboxes' to store our most cherished information, our favorite digital pictures, movies, texts and books. If we are judicious and careful, we can keep several copies of this virtual box and check from time to time that enough of them are functional. More simply, we can pay a service provider to guarantee the persistence of all this information. But, in both cases, we have to choose what we keep, and what to forget. The task is painful; the problem is too complex for us. Our only hope is that digital assistants may soon be available to help us preserve our personalized digital memory.

The opportunities and challenges of preserving digital memory are even more complex for society as a whole. To illustrate the situation, consider how the work of historians has been transformed by digitization. Once digitized, books, archives, scrolls and other documents that constitute their basic material become available anytime and anywhere. Access to our heritage becomes, in theory, technically universal. Some of the European states already digitize large sections of historical and cultural information to ensure their preservation, and make them widely accessible. Europeana,[1] the European Digital Library, launched in 2008, had more than 54 million digital objects by 2016, including text, images, and videos. With such initiatives, we can imagine that, in a few decades, historians will have at their disposal, all the information they need in digital formats, moving from one archive to another simply by changing window on their computer screen. Research will no longer have to deal with questions of distance and cost: a teacher from Abidjan will easily consult primary sources of a library in Florence.

Of course, digital connections cannot yet replace the physical contact with objects, nor the invaluable discussion with the librarian or the archivist. But digitization greatly facilitates research. For example, Optical Character Recognition can be used to transform the image of a document into a text file that can then be indexed, and analyzed. This opens countless possibilities to more easily find and access documents on a certain topic, or compare authors' writing styles.

Digitalization also provides a particular form of immortality. An old parchment will not disappear completely if the library burns; its contents will not be lost on the occasion of a move or destroyed at the whim of a tyrant; the ink will not fade over time. In this new digital age, traditional document preservation centers have thus been totally transformed. Take, for example, the Bibliothèque Nationale de France[2], established by François I in 1537 to serve as a "legal deposit", for copies of all books and official documents published in the kingdom. What happens to this concept of a physical repository of knowledge when most of the content produced is available digitally on the internet? This material is potentially 'immortal' but it may also be ephemeral in nature.

Scaling up from historical documents and archived collections, we must ask ourselves what will become of all the information of the web in fifty years, when researchers will want to study the world in which we live today? In 1994, the National Library of Canada[3] was one of the pioneers of web archiving. Then, in 1996, the Internet Archive foundation initiated a global archiving of web pages. The internet site archive.org provides access to the Wayback Machine[4]. The site is designed to go back in time, following the evolution of various versions

of internet pages. A virtual time tour through this site is both surprising and entertaining; the first incarnations of many long-established websites are deliciously simple and outdated.

Creating such internet archives requires heavy computing. Computer robots (so-called crawlers) surf the web, bringing web pages, images, videos and other web-based materials into the archive. But the web is huge, and the task daunting, even for an army of robots. To illustrate the enormity of the task, let us focus on just a single website (out of hundreds of millions). In just four years, from 2006 to 2010, the United States Library of Congress[5] recovered 170 billion tweets from a twitter.com, encompassing 133.2 terabytes of data - and this was before Twitter hosted pictures! This is just one tiny facet of the digital information we produce. We cannot archive the entire web, and it is thus necessary to focus our efforts on areas of particular interest. Typically, these efforts are directed to the most popular sites, or to the most relevant ones for particular topics such as elections. But, the web changes, new pages appear, and others evolve. We also miss considerable amounts of data that are hidden, for example, behind forms that are completed manually. These data are, therefore, out of reach of web surfers and crawlers, and access to this information requires special agreements with their producers. The effort must continue permanently, and we must accept that the version of the web we store in our evolving archive will never be complete or up to date.

Beyond archiving classical digital information such as books, pictures, and web pages, there are many other data that also need to be archived, including complex algorithms and computer programs. These products of human ingenuity and experience are part of our collective memory, our heritage, and the question of their preservation is therefore equally essential. The awareness of the software heritage preservation is recent. The Software Heritage project[6] was started in 2015 under the auspices of Inria[7]. Its purpose is to collect, store and share software codes to build a universal and freely-available archive. In just a few years, this archive has already accumulated billions of files, representing a vast trove of human ingenuity and problem solving.

Another issue for society is that the choice of information to forget should not bias what will be remembered. This archiving of the digital world should stay away from rewriting history.

The preservation and archiving of digital data both individuals and society is a quintessentially modern problem. Some of the difficulties are only temporary. Software and hardware solutions are increasingly available to facilitate the preservation of digital information and memory. But this still leaves open the question of what exactly we wish to keep. As archiving and storage become less and less expensive, the volume of data produced is exploding. This is not so much a physical problem of space to store information, as miniaturization has drastically reduced the size of digital storage devices. For example, the storage of the immense volume of data produced by CERN Large Hadron Collider weights nothing compared to that of the concrete of the 27 kilometer ring where sub-atomic particles are accelerated to near light-speed. However, a massive infrastructure is needed to support data processes centers. A decline in the cost of machinery and data storage has given us the illusion that it is possible to keep everything. But, paradoxically, our capacity to store massive quantities of information may make much of our data inaccessible in practical terms.

By entering the digital world, we have moved from a culture of relative information scarcity to one of information overload. With the continued explosion of data volume, perhaps our greatest challenge is the need to select what we want to forget. We do not have any chance of addressing this problem without the support of computer algorithms. We must learn to use

algorithms to become the archivists of our data world, striving to master a collective digital memory, which is unprecedented in the history of humanity.

---

[1] Europeana, www.europeana.eun
[2] Bibliothèque Nationale de France, www.bnf.fr
[3] National Library of Canada, gallery.ca
[4] Wayback machine, waybackmachine.org
[5] Library of Congress, loc.gov
[6] Software Heritage, softwareheritage.org
[7] Inria, inria.fr